A Computational Stochastic Modeling Formalism for Biological Networks

Werner Sandmann and Verena Wolf

Abstract — Stochastic models of biological networks are well established in systems biology, where the computational treatment of such models is often focused on the solution of the so-called chemical master equation via stochastic simulation algorithms. In contrast to this, the development of storage-efficient model representations that are directly suitable for computer implementation has received significantly less attention. Instead, a model is usually described in terms of a stochastic process or a "higher-level paradigm" with graphical representation such as e.g. a stochastic Petri net. A serious problem then arises due to the exponential growth of the model's state space which is in fact a main reason for the popularity of stochastic simulation since simulation suffers less from the state space explosion than non-simulative numerical solution techniques. In this paper we present transition class models for the representation of biological network models, a compact mathematical formalism that circumvents state space explosion. Transition class models can also serve as an interface between different higher level modeling paradigms, stochastic processes and the implementation coded in a programming language. Besides, the compact model representation provides the opportunity to apply non-simulative solution techniques thereby preserving the possible use of stochastic simulation. Illustrative examples of transition class representations are given for an enzyme-catalyzed substrate conversion and a part of the bacteriophage λ lysis/lysogeny pathway.

Keywords — Computational Modeling, Biological Networks, Stochastic Models, Markov Chains, Transition Class Models

I. Introduction

Biological network models significantly suffer from their enormous size, which is due to the high complexity and lively interactions of involved molecules. Much effort has been spent to develop and apply analysis techniques, whereas reducing the model size or, more specifically, reducing the required computer storage by providing compact formal model descriptions has received far less attention. As stated in [12], the focus of current modeling tools is on simulation, but model development is a highly iterative process which is currently only partly supported. Modelers will often end up having many different versions of one model, probably in a number of different formats.

The fundamental rule of a chemical reaction between molecules is given by the stoichiometry

$$s_{i_1}S_{i_1} + \cdots s_{i_m}S_{i_m} \longrightarrow s_{i_{m+1}}S_{i_{m+1}} + \cdots s_{i_\ell}S_{i_\ell} \quad (1)$$

with $m, \ell \in \mathbb{N}$, $m < \ell$, where $s_{i_1}, \ldots, s_{i_\ell} \in \mathbb{N}$ are *stoichiometric coefficients*, $S_{i_1}, \ldots, S_{i_m}$ are called *reactants*, $S_{i_{m+1}}, \ldots, S_{i_\ell}$ are called *products* and both reactants and products are *molecular species*. Such a chemical equation expresses that the left hand side of the arrow can be transformed to the right hand side of the arrow. Complex chemical processes are given by sets of such reactions. Although the stoichiometric coefficients specify the necessary quantities of substances of each molecular species, (1) basically describes a qualitative or functional relationship. However, for a chemical reaction to occur, typically several conditions on temperature, pressure, or concentration must hold. These are usually indicated by adding information above or below the arrow and yield quantitative and temporal relationships often given in terms of *rates*. The scientific branch that studies such rates of chemical reactions is called *chemical kinetics*.

Different types of computational mathematical models for the description of the quantitative behaviour of systems formed by chemical processes exist and the specific meaning of rates depends on the chosen model type. Though motivated by different viewpoints the model types and thus the rates are of course intimately related which should not be too surprising since they represent the same type of systems. A comprehensive treatment of different computational model types can be found in [2].

Models are distinguished in terms of their states and state changes (*transitions*) where a state consists of a collection of variables that sufficiently well represents the relevant[1] parameters of the original system at any time. The set of all states, also referred to as the *state space*, may be either discrete, meaning only a countable number of states that can be mapped to a subset of the natural numbers $\mathbb{N}$, or the state space may be continuous. Both in discrete and continuous state space models the state transitions may occur deterministically or stochastically. For a long time the model type of choice in computational systems biology was a deterministic one with continuous state space, based on the law of mass action and expressed in terms of *chemical rate equations* leading to a system of nonlinear ordinary differential equations that often turns out to be quite difficult to solve.

The stochastic approach [8], motivated by the observation that biochemical reactions occur randomly, leads to discrete state Markov processes [7], [9], [20], or, equivalently in other words, to continuous-time Markov chains [3], and it requires the solution of a system of difference-differential equations, the *chemical master equation*. The rules driving the temporal evolution of the system can be stored in a

W. Sandmann is with the Department of Information Systems and Applied Computer Science, University of Bamberg, Feldkirchenstr. 21, D-96045 Bamberg, Germany (corresponding author, phone: +49 951 863 2823; e-mail: werner.sandmann@wiai.uni-bamberg.de)

V. Wolf is with the Department of Mathematics and Computer Science, University of Mannheim, A5 Bauteil B 119, D-68131 Mannheim, Germany (email: wolf@informatik.uni-mannheim.de).

[1] Any model is a simplified abstraction of the real system and both suitability of a model and the relevant parameters depend on the scope of the study.

matrix consisting of reaction rates. The model's state space and thus the matrix dimension is determined by the number of involved molecular species and the number of potentially present molecules of each species. Unfortunately, the state space size grows exponentially in the number of molecular species and in case of potentially infinitely many molecules it is even infinite. Thus, storage of the rate matrix is not suitable for models of complex biological networks. Likewise chemical reaction equations given by the stoichiometry are not suitable for efficient modeling with regard to representation in computers.

In this paper we adopt the stochastic approach and present a structured mathematical modeling formalism called transition class model (TCM) that is particularly well suited for implementation purposes and moreover can serve as an interface between different model types or model formats. The paper is organized as follows. Section II gives formalizations of stochastic models of biological networks thereby introducing terminology and notations and demonstrating the problem of efficient storage and implementation. Transition class models are introduced in Section III, where we also outline their advantages. Section IV contains transition class representations for specific biological networks, and finally Section V concludes the paper.

II. Stochastic Modeling of Biological Networks

Stochastic interpretations of chemically reacting systems date back to the 1960s [14]. A formulation on a physical basis has been provided in [8] and later on rigorously derived in [9]. The basic assumptions are that the system is kept well stirred and thermally equilibrated, meaning that a well stirred mixture of $N \in \mathbb{N}^+$ molecular species $S_1, \ldots, S_N$ inside some fixed volume interact at constant temperature. In the following we give a brief description of the formal mathematical basis.

A. Mathematical Model Description

The system state at any time is described by a discrete random vector

$$X(t) = (X_1(t), \ldots, X_N(t)), \tag{2}$$

where for each species $S_i, i \in \{1, \ldots, N\}$ and $t \geq 0$ a discrete random variable $X_i(t)$ describes the number of S_i molecules present at time t. The conditional transient (time dependent) probability that the system is in state $x \in \mathbb{N}^N$ at time t, given that the system starts in an initial state x_0 at time t_0, is denoted by

$$p^{(t)}(x|x_0, t_0) = P\left(X(t) = x \mid X(t_0) = x_0\right). \tag{3}$$

The system state changes over time due to chemical reactions between molecules of some species. Complex reaction sets can be decomposed into elementary unidirectional reactions such that each reaction takes the form (1), where additionally a reaction rate that determines the reaction speed or probability is assigned to each reaction.

The reaction rates are independent of the time since the probability that a reaction occurs within a specific time interval only depends on the length of this interval and not on the interval endpoints (the specific start and end times). Thus, given a current system state, the next state in the system's time evolution only depends on this current system state and neither on the specific time nor on the history of reactions that led to the current state. Hence, the time evolution of the system is mathematically described by a stochastic process $(X(t))_{t\geq 0}$ with N-dimensional state space $\mathcal{S} \subseteq \mathbb{N}^N$, and due to the just stated independence of time and history this stochastic process is a discrete-state Markov process, or equivalently in other words, a time-homogeneous continuous-time Markov chain (CTMC). That is, for all $n \in \mathbb{N}$ and $t_0 < t_1 < \cdots < t_n$

$$\begin{aligned} &P\left(X(t_n) = x_n | X(t_{n-1}) = x_{n-1}, \ldots, X(t_0) = x_0\right) \\ &= P\left(X(t_n) = x_n | X(t_{n-1}) = x_{n-1}\right). \end{aligned} \tag{4}$$

The multidimensional discrete state space $\mathcal{S}$ of the CTMC can be mapped to the natural numbers $\mathbb{N}$ and the probability that a transition from state $i \in \mathbb{N}$ to state $j \in \mathbb{N}$ occurs within a time interval of length $h \geq 0$ is denoted by $p_{ij}(h)$. For all $h \geq 0$ these state transition probabilities build a transition probability matrix[2] $P(h) = (p_{ij}(h))_{ij\in\mathbb{N}}$. Note that $P(0)$ equals the unit matrix I, since no state transitions occur within a time interval of length zero.

It is well known [3], [6], [13] that a CTMC with state space $\mathcal{S} \subseteq \mathbb{N}^N$ is uniquely defined by an initial probability distribution on $\mathcal{S}$ and a *transition rate matrix*, also referred to as *infinitesimal generator matrix*, $Q = (q_{ij})_{i,j\in\mathbb{N}}$ consisting of *transition rates* q_{ij}, where

$$Q = \lim_{h\to 0} \frac{P(h) - P(0)}{h} = \lim_{h\to 0} \frac{1}{h}(P(h) - I). \tag{5}$$

The relation of each $P(h)$ to Q and an explanation for the term *infinitesimal generator matrix* is given by $P(h) = \exp(hQ)$. In that way Q generates the the transition probability matrices by a matrix exponential function which is basically defined as an infinite power series. Hence, all information on transition probabilities is covered by the single matrix Q, where in biological network modeling the transition rates q_{ij} correspond to reaction rates.

The temporal evolution of a CTMC can be described via a system of differential equations, the Kolmogorov forward equations and the Kolmogorov and backward equations, resp., in matrix notation given by

$$\frac{\partial}{\partial t}P(t) = P(t)Q, \qquad \frac{\partial}{\partial t}P(t) = QP(t), \tag{6}$$

which yields a system of differential equations for the transient state probabilities, in vector-matrix notation given by

$$\frac{\partial}{\partial t}p^{(t)} = p^{(t)}Q, \tag{7}$$

where $p^{(t)}$ denotes the vector of the transient state probabilities corresponding to (3). The above equations are equivalent to the so-called the *chemical master equation* [7], [20], a term that is

[2] A transition probability matrix is also called a stochastic matrix meaning that all entries are probabilities and all row sums equal one.

thus nothing else than a synonym for the general terms used in the theory of stochastic processes [3], [6], [13], in particular Markov processes.

Although the Kolmogorov differntial equations and the chemical master equation arise from a stochastic model there is no need to apply stochastic solution methods. In particular, there is a significant difference between a stochastic model and a stochastic simulation although in the literature ”the stochastic approach” and ”the stochastic simulation algorithm” are often taken as the same thing. In fact, as we have outlined above, the stochastic approach leads to continuous-time Markov chains that may be analyzed by a large variety of solution techniques, where stochastic simulation is only one of them.

B. Difficulties in Modeling and Analysis

Explicit algebraic solution of the Kolmogorov equations, or in the biosystems terminology of the chemical master equation, is usually impossible, and several techniques have been proposed for the numerical solution of Markov chains, see e.g. [18]. Most of these techniques both in the general context of Markov chains and in the specific application to biological systems aim to solve the above system of differential equation or a variant known as *Fokker-Planck equation*. An alternative approach to analytically cope with stochastic models of biological networks is by stochastic differential equations (SDE) that are in terms of Itô calculus (which is also very popular e.g. in stochastic finance) equivalent to the Fokker-Planck equation.

The main problem that numerical solution techniques suffer from is the enormous size of the state space that grows exponentially in the dimensionality, a problem known as *state space explosion*. In particular, for biological networks the state space grows exponentially in the number of involved molecular species, which means that even a moderate number of species implies extremely huge state spaces that are often impossible to store in computers. In case of potentially infinite molecular populations the resulting state space is even infinite. Several advanced solution techniques have been developed to deal with the state space explosion problem for specific models, most of them exploiting a special structure of the transition rate matrix and partitioning the state space to yield approximate solutions, see e.g. [4], [18], [21] and references therein. Unfortunately, if the transition rate matrix does not have the assumed structure such approximation techniques do not work.

An alternative approach that suffers less from the state space explosion problem is stochastic simulation. As already stated, the chemical master equation is equivalent to the Kolmogorov equations. Likewise, the so-called *stochastic simulation algorithm* by Gillespie [8], which is often used to solve the chemical master equation is a straightforward application of Monte Carlo simulation methods for Markov chains that are known at the latest since the early 1950s, as indicated by [6], [10], [15] and the references therein. Although often equated with the stochastic approach to modeling biochemically reacting systems the stochastic simulation algorithm is only one specific solution technique, and its popularity is mainly justified by the difficulty of solving the differential equations with other techniques. Nevertheless, stochastic simulation has numerous drawbacks, and in many application areas where stochastic models are used, stochastic simulation is often even referred to as a method of last resort.

One of the major drawbacks of stochastic simulation is the random nature of simulation results. Despite the fact that Gillespie’s algorithm is termed exact, a stochastic simulation can never be exact. Mathematically, it constitutes a statistical estimation procedure implying that the results are subject to statistical uncertainty and in order to draw meaningful conclusions it is necessary to make statistically valid statements on the results. The exactness of Gillespie’s algorithm is only ”in the sense that it takes full account of the fluctuations and correlations” [8] of reactions within a single simulation run and Gillespie mentions that it is ”necessary to make several simulation runs from time 0 to the chosen time t, all identical with each other except for the initialization of the random number generator”. In fact the reliability of simulation results strongly depends on a sufficiently large number of simulation runs, and a proper determination of that number has to be carefully done in terms of mathematical statistics.

Furthermore, stochastic simulation is inherently costly. In many cases even a single simulation run is extremely computer time demanding and thus reducing the space complexity compared to numerical methods has to be paid by a significant increase of time complexity. Serious difficulties arise in the presence of multiple time scales or stiffness. Often approximations are required to achieve simulation speed up, and as an immediate consequence even the exactness in the sense stated above gets lost. Thus, if a problem may be tackled both by stochastic simulation and by numerical analysis, the latter should be preferred. The difficulties in numerical analysis mainly arise due to the state space explosion. Hence, it is highly desirable to develop compact modeling formalisms that render model representation and storage in a computer possible and that yield to numerical analysis as well.

III. Transition Class Models

To avoid the problem of state space explosion, we use transition class models (TCMs), which are compact and structured formal descriptions of Markov chains. They are originally motivated by queueing network state spaces and similarity of state transitions in this context, but it turns out that they are also well suited for formalizing biochemically reacting systems. It is not necessary, but possible, to generate the complete state space and the transition rate matrix explicitly. Once a TCM has been developed, many different solution techniques, including stochastic simulation, can be applied.

Algorithms have been developed to generate transition class models automatically from formal Petri net and queueing network descriptions [17], [19]. Hence, TCMs have the potential to serve as an interface between different model specifications

(queueing models, Petri nets, mixtures of them, amongst many others) and various solution methods. Different parts of a model can be described by different modeling paradigms that may be on different levels of abstraction, e.g. parts are given as queueing model, other parts as Petri net, some parts may be specified as a Markov chain on the low abstraction level of a stochastic process, others via structured stochastic automata networks as recently done for biochemically reacting systems in [21]. Transformation into a TCM then yields a unified model description, which is moreover suitable for immediate solution.

Fig. 1 illustrates how transition class models are integrated within the development of an implementation for a system description, in particular showing their interface character.

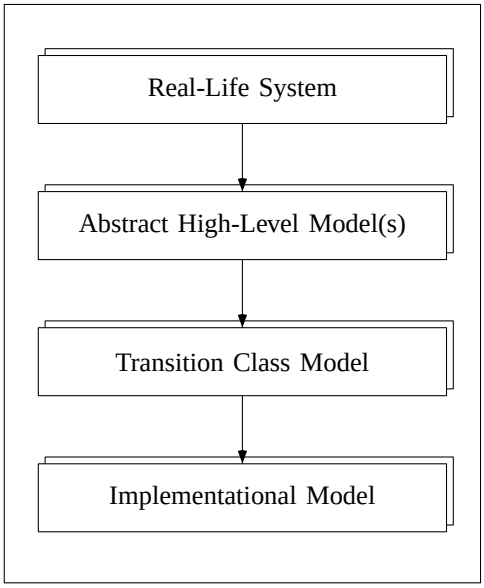

Fig. 1. Integration of Transition Class Models within the "Modeling to Implementation" Process

A. Formal Definitions and Properties

Although in systems biology the interest is usually in transient state probabilities there are also relevant cases where steady-state probabilities – probabilities for a system in equilibrium – provide important insights and are thus of interest. It is well known from the theory of stochastic processes [3], [6], [7], [13], [18], [20] that steady-state probabilities for continuous-time Markov chains can be derived via an embedded discrete-time Markov chains, where state transitions occur only after discrete time steps according to transition probabilities, which is sometimes easier to analyze (depending on the chosen solution technique). Accordingly, we provide definitions of transition class models for both the continuous-time case and the discrete-time case. We follow the presentation in [16], where a formal definition of transition class models appeared for the first time. Essentially for the structured description is the notion of a transition class, which enables us to interpret and model state transition events efficiently, e.g. reactions in biological networks.

Definition 1: (Transition Class) A transition class (relative to some set $\mathcal{S}$) is a triplet $\tau = (\mathcal{U}, u, \alpha)$ consisting of

- a set $\mathcal{U}$,
- a function $u : \mathcal{U} \cap \mathcal{S} \to \mathcal{S}$, where $\forall x \in \mathcal{U} \cap \mathcal{S} : u(x) \neq x$,
- $\begin{cases} \text{a function } \alpha : \mathcal{U} \cap \mathcal{S} \to (0,1] & \text{in discrete time,} \\ \text{a function } \alpha : \mathcal{U} \cap \mathcal{S} \to (0,\infty) & \text{in continuous time.} \end{cases}$

For $\alpha : \mathcal{U} \cap \mathcal{S} \to (0,1]$ we speak of a discrete transition class (DTC), and for $\alpha : \mathcal{U} \cap \mathcal{S} \to (0,1]$ we speak of a continuous transition class (CTC).

Next we give an interpretation for what is described by a transition class, and we introduce an appropriate terminology.

The set $\mathcal{U}$ contains states, e.g. describing a system represented by a model. These states may change when some events (state transitions) occur. Therefore, we refer to $\mathcal{U}$ as the *source state space* of τ. Note that we allow $\mathcal{U} \setminus \mathcal{S} \neq \emptyset$, which means, $\mathcal{U}$ may contain some redundant (infeasible) states. This makes formal model description much easier and more efficient. Additionally, we emphasize that we need not explicitly specify the set $\mathcal{S}$ when defining concrete transition classes, and neither all elements of the source state space have to be enumerated nor have they to be stored completely.

The function u gives the new state after a transition from one state to another state (which need not be contained in $\mathcal{U}$) has occured. Therefore, we call u the *destination state function* (or *target state function*, which is more familiar in some areas). Note that from the definition of the destination state function it immediately follows that the source state space of any transition class does not contain absorbing states, i.e. states where the system stays forever if once reached. In the discrete case the definition additionally implies that state transitions from a state to itself, so called self–loops, corresponding to positive diagonal entries in the transition probability matrix when using classical Markov chain descriptions, need not to be modeled explicitly as a transition class. Thus, an additional source of storage waste is eliminated.

Finally, $\alpha(x)$ denotes for a DTC the probability and for a CTC the rate of such a transition from state x to state $u(x)$. For DTC we call α the *transition probability function*, and for CTC we call α the *transition rate function*. We point out, that in many cases, when transition classes are defined properly, α is a constant, i.e. it does not depend on the system state, or at the worst it is a rather simple function on the system state.

Now we are ready to give the formal definition of a transition class model, both for discrete and continuous time.

Definition 2: (Transition Class Model, TCM)
Let $\mathcal{T} := (\{\tau_1, \ldots, \tau_k\}, y)$ be a pair consisting of a set of transition classes $\tau_i = (\mathcal{U}_i, u_i, \alpha_i), 1 \leq i \leq k$ and a feasible state $y \in \mathcal{S} \cap (\mathcal{U}_1 \cup \ldots \cup \mathcal{U}_k)$. Then $\mathcal{T}$ is called a *continuous transition class model* (CTCM), if each τ_i is a CTC; and $\mathcal{T}$ is called a *discrete transition class model* (DTCM), if each τ_i is a DTC, and

$$\forall x \in \mathcal{S} \cap \bigcup_{i=1}^{k} \mathcal{U}_i : \sum_{i=1}^{k} I_{\{x \in \mathcal{U}_i\}} \alpha_i(x) \leq 1. \tag{8}$$

If in inequality (8) for states x the condition "< 1" holds, then there is a positive probability of a self–loop in x, and this probability is exactly the difference to one. As we have stated earlier, self–loops are not modeled explicitly, but are implicitly contained in the transition class model.

What has been gained compared to the usual Markov chain description via a transition rate matrix? Typically, the Markov chain state space grows exponentially, whereas the number of transition classes grows only linearly in the number of molecular species. Moreover, it is possible to describe Markov chains with infinite state space by a finite number of transition classes. Consider for example a potentially infinite number of at least one of the involved molecular species. Then the source state spaces of course become infinite, but they can still be described by component characteristics meaning by characteristics of single molecular species.

Intuitively, it seems clear, that Markov chains can be described as transition class models, and indeed it can be formally proven, that each Markov chain can be described by a TCM, and that each TCM can be interpreted as and thus describes a Markov chain [16]. Formally, a transition class model is an abstract mathematical notation, which gets a practical meaning and a relation to other modeling paradigms only by interpreting its components. The interpretation as a Markov chain thus yields in this sense a semantics of transition class models. Each τ_i is a transition class relative to some set $\mathcal{S}$ without $\mathcal{S}$ explicitly given in the definition. This means, that a TCM implicitly contains the state space of the described Markov chain. In particular, using TCMs does neither require any numbering of states nor explicit enumeration of the state space, and TCM can be stored very efficiently.

Obviously, TCMs both in continuous and in discrete time can be simulated in a similar manner as Markov chains by repeatedly generating trajectories, as e.g. in its easiest and most straightforward way adopted by Gillespie in his stochastic simulation algorithm [8]. Again note that [8] is by no means the first paper where direct Markov chain simulation appears. Moreover, although not specifically concerned with simulation, TCMs are well-suited for improved fast simulation methods that are far more advanced than the Gillespie algorithm and its variants, for instance variance reduction techniques based on importance sampling [16]. Even more important with transition class models there is no need to resort to stochastic simulation since non-simulative numerical techniques can be directly performed on TCMs. Hence, as a natural by-product of circumventing the problem of state space explosion, transition class models open access to a much wider range of analysis methodologies.

IV. Transition Class Representations

We demonstrate how transition class representations of concrete biological networks look like by illustrating it via example for an enzyme-catalyzed substrate conversion and a part of the bacteriophage λ lysis/lysogeny pathway.

A. Enzyme-catalyzed Substrate Conversion

As the first example consider a representative system that has been also served as a reference example, e.g. very recently in [4], [5], the enzyme-catalyzed substrate conversion

$$S_1 + S_2 \underset{c_2}{\overset{c_1}{\rightleftharpoons}} S_3 \xrightarrow{c_3} S_1 + S_4 \tag{9}$$

of a *substrate* S_2 into a *product* S_4 via an *enzyme-substrate complex* S_3, *catalyzed* (accelerated) by an enzyme S_1.

If we assume that initially (at time 0) there are $x_1^{(0)}$ enzyme molecules, $x_2^{(0)}$ substrate molecules, and no molecules of the enzyme-substrate complex and the product are present, then the maximum numbers of molecules of S_1 and S_3 that can be present at any time t are $x_1^{(0)}$, and for S_2 and S_4 they are $x_2^{(0)}$. Hence, the state space size of the corresponding Markov chain equals $|\mathcal{S}| = (x_1^{(0)} + 1) \cdot (x_2^{(0)} + 1)$ which yields e.g. for $x_1^{(0)} = 200$ and $x_2^{(0)} = 3000$ the size of $201 \cdot 3001 \approx 6 \cdot 10^5$. If we do not have bounds for the initial molecule population it is infinite. In our representation we need only three transition classes τ_1, τ_2, τ_3 even in case of an infinite state space:

$\tau_1 = (\mathcal{U}_1, u_1, \alpha_1)$, where
- $\mathcal{U}_1 = \{(x_1, \ldots, x_4) : x_1, x_2 > 0\}$,
- $u_1 : \mathbb{N}^4 \to \mathbb{N}^4$,
  $x \mapsto u_1(x) = (x_1 - 1, x_2 - 1, x_3 + 1, x_4)$,
- $\alpha_1 : \mathbb{N}^4 \to \mathbb{R}, \quad x \mapsto \alpha_1(x) = c_1 x_1 x_2$;

$\tau_2 = (\mathcal{U}_2, u_2, \alpha_2)$, where
- $\mathcal{U}_2 = \{(x_1, \ldots, x_4) : x_3 > 0\}$,
- $u_2 : \mathbb{N}^4 \to \mathbb{N}^4$,
  $x \mapsto u_2(x) = (x_1 + 1, x_2 + 1, x_3 - 1, x_4)$,
- $\alpha_2 : \mathbb{N}^4 \to \mathbb{R}, \quad x \mapsto \alpha_2(x) = c_2 x_3$;

$\tau_3 = (\mathcal{U}_3, u_3, \alpha_3)$, where
- $\mathcal{U}_3 = \mathcal{U}_2 = \{(x_1, \ldots, x_4) : x_3 > 0\}$,
- $u_3 : \mathbb{N}^4 \to \mathbb{N}^4$,
  $x \mapsto u_3(x) = (x_1 + 1, x_2, x_3 - 1, x_4 + 1)$,
- $\alpha_3 : \mathbb{N}^4 \to \mathbb{R}, \quad x \mapsto \alpha_3(x) = c_3 x_3$;

Obviously, the TCM provides a huge gain in storage requirements and is well suited for immediate implementation. An important point regarding computer implementations is that the state space and the transition rate matrix of the underlying Markov chain is implicitly coded by logical predicates and simple functions that are both easy to implement.

B. Lambda Bacteriophage

In this example we develop a TCM model for a part of the bacteriophage λ lysis/ lysogeny pathway. We focus on the $\mathsf{P_R} - \mathsf{P_{RM}}$ operator regions sharing several overlapping operator sites. The expression of the λ repressor gene cI is a well characterized autoregulated genetic network (see [1], [11] and the references therein). The mutant system has operator sites OR2 and OR3 where the CI dimer[3], denoted by X_2, binds as a transcription factor either 1) at OR2, 2) at OR3 or

[3]Here, gene names start with lower case letters and the corresponding proteins are denoted by upper case letters

3) at both sites. In case 1, i.e. X_2 binds at OR2, transcription is enhanced whereas binding at OR3 (cases 2 and 3) inhibits transcription which means that the production of protein CI is turned off. Let D denote the DNA promotor site. The stoichiometry of the model is given in Table I and since there are 13 different reaction types and 6 species its TCM model requires 13 transition classes and $\mathcal{S} = \mathbb{N}^6$. We assume that in state $x = (x_1, x_2, \ldots, x_6)$ the population sizes of the 6 species $X, X_2, D, DX_2, DX_2^*, DX_2X_2$ are x_1 for X, x_2 for $X_2, \ldots,$ and x_6 for DX_2X_2.

TABLE I

STOICHIOMETRY OF THE LYSIS-LYSOGENY SWITCH IN BACTERIOPHAGE λ.

$2X$	$\overset{c_1}{\rightleftharpoons}$	X_2	dimerization
$D + X_2$	$\overset{c_2}{\rightleftharpoons}$	DX_2	binding 1)
$D + X_2$	$\overset{c_3}{\rightleftharpoons}$	DX_2^*	binding 2)
$DX_2 + X_2$	$\overset{c_4}{\rightleftharpoons}$	DX_2X_2	binding 3)
$DX_2^* + X_2$	$\overset{c_5}{\rightleftharpoons}$	DX_2X_2	binding 3)
D	$\xrightarrow{c_s}$	$D + X$	slow transcription
X	$\xrightarrow{c_d}$	$\emptyset$	degradation
DX_2	$\xrightarrow{c_f}$	$DX_2 + X$	enhanced transcription

Then, for instance, reaction $2X \xrightarrow{c_1} X_2$ is described by transition class $\tau_1 = (\mathcal{U}_1, u_1, \alpha_1)$ where

- $\mathcal{U}_1 = \{(x_1, x_2, \ldots, x_6) : x_1 \geq 2\}$,
- $u_1 : \mathbb{N}^6 \to \mathbb{N}^6$,
  $x \mapsto u_1(x) = (x_1 - 2, x_2 + 1, x_3, x_4, x_5, x_6)$
- $\alpha_1 : \mathbb{N}^6 \to \mathbb{R}$,
  $x \mapsto \alpha_1(x) = 2c_1x_1$.

Since the population of D is at most one, the transition class of reaction $D + X_2 \xrightarrow{c_2} DX_2$ is $\tau_2 = (\mathcal{U}_2, u_2, \alpha_2)$ where

- $\mathcal{U}_2 = \{(x_1, x_2, \ldots, x_6) : x_2 > 0, x_3 = 1\}$,
- $u_2 : \mathbb{N}^6 \to \mathbb{N}^6$,
  $x \mapsto u_2(x) = (x_1, x_2 - 1, 0, x_4 + 1, x_5, x_6)$
- $\alpha_2 : \mathbb{N}^6 \to \mathbb{R}$,
  $x \mapsto \alpha_2(x) = c_2x_2$.

Reaction $DX_2X_2 \xrightarrow{c_5} DX_2^* + X_2$ is described by transition class $\tau_3 = (\mathcal{U}_3, u_3, \alpha_3)$ where

- $\mathcal{U}_3 = \{(x_1, x_2, \ldots, x_6) : x_6 = 1\}$,
- $u_3 : \mathbb{N}^6 \to \mathbb{N}^6$,
  $x \mapsto u_3(x) = (x_1, x_2 + 1, x_3, x_4, x_5 + 1, 0)$
- $\alpha_3 : \mathbb{N}^6 \to \mathbb{R}$,
  $x \mapsto \alpha_3(x) = c_5$.

The three transition classes given above should suffice for illustration. The remaining ones are built in the same manner. Again, TCMs provide a huge gain in required computer storage. It becomes clear that this gain rapidly increases with the model size, because for a Markov chain description the model size grows exponentially in the number of involved molecular species whereas the size of a transition class model grows only linearly.

V. CONCLUSIONS

We have presented transition class models as a mathematical formalism for the compact and structured representation of stochastic models for biological networks. Transition class models provide huge gains in computer storage requirements, are well suited for implementation and may also serve as an interface between different high level modeling paradigms. Moreover, they open access to a wide range of analysis methodologies that are not feasible when using classical Markov process descriptions. Transition class representations have been illustrated for an enzyme-catalyzed substrate conversion and a part of the bacteriophage λ lysis/lysogeny pathway. Ongoing research is concerned with improving and extending already existing numerical solution techniques that directly work with the transition class representation and also with advanced stochastic simulation algorithms for transition class models.

REFERENCES

[1] A. Arkin, J. Ross, and H. H. McAdams, "Stochastic Kinetic Analysis of Developmental Pathway Bifurcation in Phage λ-Infected Escherichia coli Cells," *Genetics*, vol. 149, pp. 1633–1648, 1998.

[2] J. M. Bower and H. Bolouri (eds.), *Computational Modeling of Genetic and Biochemical Networks*. Cambridge, MA: The MIT Press, 2001.

[3] P. Bremaud, *Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues*. Berlin–Heidelberg–New York–Tokyo: Springer–Verlag, 1999.

[4] H. Busch, W. Sandmann, and V. Wolf, "A Numerical Aggregation Algorithm for the Enzyme-catalyzed Substrate Conversion," in *Proc. Int. Conf. on Computational Methods in Systems Biology (CMSB)*, Trento, Italy, 18-19 October, to appear, 2006.

[5] Y. Cao, D. T. Gillespie and L. R. Petzold, "Accelerated Stochastic Simulation of the Stiff Enzyme-Substrate Reaction," *J. Chemical Physics*, vol. 123, no. 14, 2005.

[6] D. R. Cox and H. D. Miller, *The Theory of Stochastic Processes*. London: Chapman and Hall, 1965.

[7] C. W. Gardiner, *Handbook of Stochastic Methods for Physics, Chemistry and the Natural Sciences*. Berlin–Heidelberg–New York–Tokyo: Springer–Verlag, 3rd ed., 2004.

[8] D. T. Gillespie, "Exact Stochastic Simulation of Coupled Chemical Reactions," *J. Physical Chemistry*, vol. 81, no. 25, pp. 2340–2361, 1977.

[9] D. T. Gillespie, "A Rigorous Derivation of the Chemical Master Equation," *Physica A*, vol. 188, pp. 404–425, 1992.

[10] J. M. Hammersley and D. C. Handscomb, *Monte Carlo Methods*. London: Methuen, 1964.

[11] J. Hasty, J. Pradines, M. Dolnik, and J. J. Collins, Noise-based switches and amplifiers for gene expression. *Proc. Natl. Acad. Sci. USA*, vol. 97, no. 5, pp. 2075–80, 2000.

[12] B. H. Junker, D. Koschützki and F. Schreiber, "Kinetic Modelling with the Systems Biology Modelling Environment SyBME," *J. Integrative Bioinformatics*, 0018, 2006. *Online Journal:* http://journal.imbio.de/index.php?paper_id=18

[13] S. Karlin and H. M. Taylor, *A First Course in Stochastic Processes*. New York: Academic Press, 2nd ed., 1975.

[14] D. A. McQuarrie, "Stochastic Approach to Chemical Kinetics," *J. Applied Probability*, vol. 4, pp. 413–478, 1967.

[15] H. A. Meyer (ed.), *Symposium on Monte Carlo Methods*. New York: John Wiley & Sons, 1954.

[16] W. Sandmann, "Importance Sampling for Transition Class Models," in *Proc. 3rd Int. Workshop on Rare Event Simulation*, Pisa, Italy, 2000.

[17] G. Siersetzki, "Algorithmen zur Erzeugung von Übergangsklassen–Modellen aus erweiterten stochastischen Petrinetzen (in German)," Master's Thesis, Universität Bonn, Institut für Informatik, 2000.

[18] W. J. Stewart, *Introduction to the Numerical Solution of Markov Chains*. Princeton University Press, 1994.

[19] J. C. Strelen, "Generation of Transition Class Models from Formal Queueing Network Descriptions," in *Proc. 12th European Simulation Symposium*, pp. 525–530, 2000.

[20] N. G. Van Kampen, *Stochastic Processes in Physics and Chemistry*. North-Holland: Elsevier, 1992.

[21] V. Wolf, "Modelling of Biochemical Reactions by Stochastic Automata Networks," *Proc. Workshop on Membrane Computing and Biologically Inspired Process Calculi (MeCBIC)*, Venice, Italy, July 9, 2006.